

# 基于粒子群算法的数据库查询优化

林桂亚

(浙江警官职业学院, 杭州 310018)

**摘 要:** 研究粒子群算法在数据库查询优化中的应用问题。为了解决大型数据库信息检索困难、查询效率低的问题,提出了一种基于粒子群算法优化数据库查询技术方案。算法提出了一种数据库查询执行计划代价模型,主要包括了查询多链接次序以及副本的选择问题,准确定义了数据库查询执行代价,采用提出的粒子群算法来优化并求解该执行代价问题,从而使得分组数目更少、数据定位更精确。实例验证结果表明,通过属性表现和违规行为任何教师都可以被准确定位,减少了分组,为数据库查询提供了优化。

**关键词:** 查询优化; 粒子群算法; 数据库查询优化; 分组查询

**中图分类号:** TP311

**文献标志码:** A

**文章编号:** 1001-3695(2012)03-0947-03

doi:10.3969/j.issn.1001-3695.2012.03.041

## Optimization of database query application research based on particle swarm algorithm

LIN Gui-ya

(Zhejiang Police Vocational Academy, Hangzhou 310018, China)

**Abstract:** This paper studied on particle swarm algorithm in database query optimization problems in application. In order to solve large-scale database information retrieval difficult, query efficiency low at the end of the problem, this paper proposed one kind based on the particle swarm optimization algorithm for database query technology. The algorithm mainly proposed a database query execution plan cost model, including the query multiple link order and replica selection problem, accurate definition of database query execution costs, then using the particle swarm algorithm to optimize the execution cost and solving problems, so that the data packet being fewer in number and more accurate positioning. The final test results show that, verified any tea-chers through attribute and irregularities can be accurate positioning, reduces the packet, for the database query provides optimization.

**Key words:** query optimization; particle swarm optimization(PSO); database query optimization; query packet

## 0 引言

数据库查询优化问题一直是国内外研究热点。为了满足数据库中信息检索需求,如何从这些信息中发掘出真正需要的信息成为当今世界人们面临的一个普遍难题,这就需要研究新的技术手段来组织、整理和存取海量信息。信息检索<sup>[1]</sup>致力于寻求准确、高效、智能的信息组织与存取方法,由此成为当前计算机相关学科的一个研究热点。之后的一些学者试图将很多算法应用到数据库中,其中的一项技术是使用模糊集在数据库中检索有意义的完整数据。从大型数据库中检索信息是不容易的,但这些被检索出的有用信息的作用往往十分关键。通过引入等价类和等价关系<sup>[2]</sup>可以降低数据库的复杂度,而在数据库属性上使用模糊集<sup>[3,4]</sup>却能进一步简化获取所需信息的方法以及拓宽查询的范围。

传统的数据库查询优化模型包括布尔模型、向量空间模型和概率模型<sup>[2]</sup>。布尔模型中文档与查询的匹配过于严格,对检索词的重要程度无法区分,无法进行检索结果的排序。为了解决大型数据库信息检索困难、查询效率低的问题,本文提出了一种基于粒子群算法优化数据库查询技术方案。算法提出了一种数据库查询执行计划代价模型,主要包括了查询多链接

次序以及副本的选择问题,准确定义了数据库查询执行代价,然后采用本文提出的粒子群算法来优化并求解该执行代价问题,从而使得分组数目更少、数据定位更精确。最后的实例验证结果表明,通过属性表现和违规行为,任何教师都可以被准确定位,减少了分组,为数据库查询提供了优化。

## 1 数据库查询执行代价模型建立

数据库的查询计划可以用查询二叉树  $p(v, e)$  来表示。其中:  $v$  表示查询二叉树的节点,是查询的关系序列;  $e$  表示查询二叉树的边,主要是连接操作。叶子节点主要表示基本的关系,而对于同一个数据库查询的语句,由于关系副本的选择,关系连接的顺序等诸多方面都存在不同,所以有不同的查询执行计划。

本文主要考虑了数据库的查询时间以及网络的传输性能优化两个方面的因素,主要选择最优的查询执行计划。数据库查询的时间主要是指查询树中查询代价最大的节点分支。语句查询代价的数学表达式为

$$\cos t(p) = \sum_{v \in V} \cos t(v) \quad (1)$$

其中:  $\cos t(v)$  表示每一个节点数据库查询执行代价,  $v$  表示二叉树节点。

## 2 粒子群算法优化代价函数

### 2.1 粒子群算法

粒子群优化 PSO 算法是一种基于群智能的随机全局优化计算技术,其优势在于算法的简洁性,易于实现,没有很多参数需要调整,且不需要梯度信息。根据粒子群的基本收敛性质,Sun 等人提出了一种基于量子空间的粒子群算法,是对标准 PSO 算法的全新改进,大量实验验证了 QDPSO 算法可提高粒子群算法的全局收敛能力。

在量子空间中,粒子的速度和位置是不能同时确定的,因此,粒子的状态必须用波函数  $\Psi(\boldsymbol{x},t)$  来描述,其物理意义是:波函数模的平方是粒子在空间某一点出现的概率密度,即

$$|\Psi|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z = Q\mathrm{d}x\mathrm{d}y\mathrm{d}z \tag{2}$$

其中: $Q$  表示概率密度函数。

这个概率分布密度函数也是满足归一化条件的。

$$\int_{-\infty}^{+\infty} |\Psi|^2 \mathrm{d}x\mathrm{d}y\mathrm{d}z = \int_{-\infty}^{+\infty} Q\mathrm{d}x\mathrm{d}y\mathrm{d}z = 1 \tag{3}$$

在量子世界中粒子运动的动力学方程为

$$i\hbar \frac{\partial}{\partial t} \Psi(\boldsymbol{x},t) = H\Psi(\boldsymbol{x},t) \tag{4}$$

其中: $H$  表示哈密顿算子; $\hbar$  为普朗克常数。

主要算法可以描述如下:

$D$  维搜索空间中,有  $m$  个粒子,其中第  $i$  个粒子的位置是  $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$ , 记第  $i$  个粒子搜索到的最优位置为  $\boldsymbol{p}_i = (p_{i1}, p_{i2}, \cdots, p_{iD})$ , 整个粒子群搜索到的最优位置为  $\boldsymbol{p}_g = (p_{g1}, p_{g2}, \cdots, p_{gD})$ , 势中心点  $\boldsymbol{p} = (p_1, p_2, \cdots, p_D)$ , 其中  $p_d = \frac{(r_1 p_{id} + r_2 p_{gd})}{(r_1 + r_2)}$ 。

粒子状态更新操作为

$$L = z \cdot |x_{id}(t) - p| \tag{5}$$

如果当  $u > 0.5$ , 有

$$x_{id}(t+1) = p - L \cdot \ln\left(\frac{1}{u}\right) \tag{6}$$

否则,有

$$x_{id}(t+1) = p + L \cdot \ln\left(\frac{1}{u}\right) \tag{7}$$

其中: $i = 1, 2, \cdots, m; d = 1, 2, \cdots, D; z$  是小于  $\frac{1}{\ln 2}$  的非负常数; $u, r_1$  和  $r_2$  是介于  $[0, 1]$  之间的随机数; $t$  为当前迭代次数。

### 2.2 编码方案选择

本文采用的是串结构的编码方案,其中粒子群的染色体分成了链接的次序以及等价关系和关系副本选择的编码段三个部分。对于某一个数据库查询语句,通过查询全局数据字典来获得其之间的等价关系,并作为其之间的启发信息,同时生成关系副本部分的编码。

对于操作节点以及连接顺序之间的编码,本文采用了属性结构转换成的串结构的编码方式,将连接操作以及等价关系依据查询树的先根遍历访问来进行编码操作。本文采用证书来表示关系叶子节点,采用大写的字母来表示关系中间节点。

### 2.3 适应度函数

数据库查询优化的主要目的就是使得最优代价执行计划达到最小。因为粒子群算法在生成新的种群时,其主要是选择概率比较大的,这就需要设定适应度函数,定义如下:

$$f(\text{pop}, (\text{tree})) = 1/\cos t(\text{pop}(\text{tree})) \tag{8}$$

一般来说,适应度函数值越大,对应的查询执行代价就越小,这也说明了查询执行的时间少,查询计划就好。

## 3 实验结果与分析

为了验证本文算法的有效性,通过从一个大数据库中查询取得所需信息来进行实验,表 1 为示例数据库。

表 1 教师信息

| 教师编号 | 科目   | 章节  | 表现 | 违规行为 |
|------|------|-----|----|------|
| 1    | 操作系统 | 章 A | 中等 | 10   |
| 2    | 操作系统 | 章 B | 良好 | 7    |
| 3    | 多媒体  | 章 A | 优秀 | 10   |
| 4    | 操作系统 | 章 C | 中等 | 9    |
| 5    | 多媒体  | 章 B | 中等 | 5    |
| 6    | 数据库  | 章 A | 差  | 11   |
| 7    | 操作系统 | 章 D | 中等 | 10   |
| 8    | 多媒体  | 章 C | 良好 | 6    |
| 9    | 操作系统 | 章 A | 优秀 | 4    |
| 10   | 数据库  | 章 B | 优秀 | 6    |
| 11   | 操作系统 | 章 F | 良好 | 9    |
| 12   | 数据库  | 章 C | 优秀 | 8    |
| 13   | 多媒体  | 章 D | 优秀 | 10   |
| 14   | 数据库  | 章 D | 良好 | 4    |

首先指定给出的数据库不同属性分别为教师编号、科目、分配给每个教师的章节、每位教师在科目中的表现以及违规行为,违规行为反映了教师上课时的违规行为所占的比重。

类(表现) = {优秀, 良好, 中等, 差}

可以认为优秀和良好与较好的表现是等价的。相似地,中等和差与较差的表现是等价的。假设如下条件成立:一个教师可以讲授多个科目;一个科目可以分配到多个教师,但是必须对不同的教师分配不同的章节,也就是说,在同一科目必须有多个章节,如章 A 和章 B。假如这门科目是操作系统,就可以把其中的章节 A 分配给教师 X 而把章节 B 分配给教师 Y。

假设一个用户只是想知道成绩是较好还是较差,属性“表现”的域有四个值:

$$D_{\text{per}} = \{\text{优秀, 良好, 中等, 差}\}$$

通过使用等价关系,得到的  $E$  如表 2 所示。

表 2 等价关系

| $E$ | 优秀 | 良好 | 中等 | 差 |
|-----|----|----|----|---|
| 优秀  | 1  | 1  |    |   |
| 好   | 1  | 1  |    |   |
| 中等  |    |    | 1  | 1 |
| 差   |    |    | 1  | 1 |

本文将教师数据库进行分组,以方便找出那些课堂表现不好的教师,并对他们采取相应的措施使他们适应组织。对属性“表现”和“违规行为”的域定义为

$$D_{\text{per}} = \{\text{优秀, 良好, 中等, 差}\}$$

$$D_{\text{irr}} = \{1, 12\}$$

通过以下观点来定义查询:优秀和良好是较好的表现;中等和差是较差的表现。

$$G_{\text{per}} = \{\{\text{优秀, 良好}\}, \{\text{中等, 差}\}\}$$

同样的,违规行为在  $\{1, 7\}$  范围内是可以接受的;违规行

为在 [8,14] 范围内是不可接受的。

$$G_{irr} = \{ \{1,7\}, \{8,14\} \}$$

基于如上观点,定义四组数据库如表 3 所示。

表 3 定义组

| 教师              | 表现        | 违规行为        |
|-----------------|-----------|-------------|
| { 2, 9,11, 12 } | { 优秀,良好 } | { 4,6,7 }   |
| { 9, 11, 12 }   | { 优秀,良好 } | { 8,10 }    |
| { 4 }           | { 中等 }    | { 5 }       |
| { 1, 4, 5, 6 }  | { 中等,差 }  | { 9,10,11 } |

通过减少组得到的矩阵如图 1 所示。从图 1 中可知,根据每个教师所属的组就能知道应对他们采取什么样的行动。但此矩阵存在一个问题,即使教师提高了自己的表现,并在一定程度上减少了违规行为,他们还是在同一类中。为了解决这个问题,可采用具有属性表现和违规行为的二叉树集。

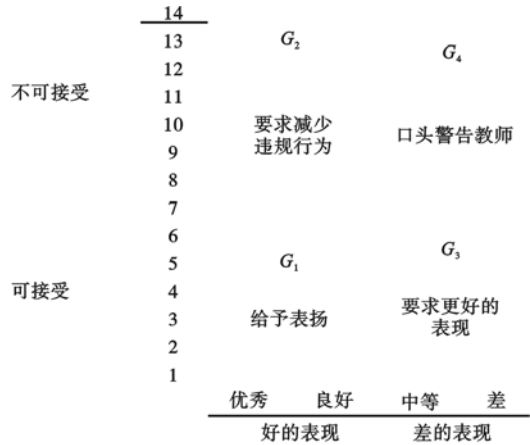


图1 矩阵

4 结束语

数据库查询优化问题比较复杂,数据冗余存储的增加以及各个节点性能的不同,使得数据库查询复杂费时。本文为了降低大数据库的复杂性,提出了一种基于粒子群算法优化数据库

查询技术方案。算法提出了一种数据库查询执行计划代价模型,主要包括了查询多链接次序以及副本的选择问题,准确定义了数据库查询执行代价,然后采用本文提出的粒子群算法来优化并求解该执行代价问题,从而使得分组数目更少,数据定位更精确。最后的实例验证结果表明,通过属性表现和违规行为为任何教师都可以被准确定位,减少了分组,为数据库查询提供了优化。

参考文献:

[1] ĆIRIĆ M, IGNIJATOVIĆ J, BOGDANOVIĆ S. Fuzzy equivalence relations and their equivalence classes[J]. Fuzzy Sets and Systems, 2007,158(12):1295-1313.

[2] BONEH D, GENTRY C, WATERS B. Collusion resistant broadcast encryption with short ciphertexts and private keys[C]//Proc of Advances in Cryptology-Crypto. 2005:258- 275.

[3] 蔡永泉,张雪迪,姜楠. 一种新的基于身份的门限签名方案[J]. 电子学报, 2009(S1):102- 105.

[4] GALINDO J, PIATTNI M, URRUTIA A. Fuzzy database modeling design & implementation[M]. Pennsylvania:IGI,2005.

[5] SONG B S, LEE K M, LEE S U. Model-based object recognition using geometric invariants of oints and lines[J]. Computer Vision and Image Understanding,2001,84(3):361-381.

[6] BARKAOUI K, BENAMARA R. On concurrency control in multidatabase systems with an extended transaction model[J]. Journal of Supercomputing,2003,24(2):193-202.

[7] 肖卫军,卢正鼎,李兵,等. 一种多数据库事务模型[J]. 小型微型计算机系统, 2008,24(12):2224-2227.

[8] WEIKUM G, VOSSEN G. Transactional information systems: theory, algorithms, and the practice of concurency control and recovery [M]. San Francisco:Morgan Kaufmann,2011.

[9] 易和平. 分布式多数据库高校学籍管理系统研究与应用[J]. 西安石油大学学报, 2009,24(4):93-95.

[10] 张敏,冯登国,徐震. 多级多版本数据库管理系统全局串行化(英文)[J]. 软件学报, 2007,18(2):346-347.

(上接第 943 页)

$$M_1 = M_0 + S_1 \times I \times U_1 = (100100000000000000) \quad (7)$$

相应的  $R_2 = \{ P_4 \}$ , 根据故障传播路径计算公式  $R = R + R_j$ ,此时  $R = \{ P_1, P_4 \}$ 。如此依次向下推导,得到最终的状态  $M_3$  为

$$M_3 = M_2 + S_3 \times I \times U_3 = (100100001000000010) \quad (8)$$

此时目标库所  $P_{17}$  所代表的接线端子接线不良故障为底故障,将  $P_{17}$  加入到可能故障发生原因列表中,并将其故障传播路径  $R = \{ P_1, P_4, P_9, P_{17} \}$  保存在数据结构中。

故障发生原因具有多样性和复杂性,因此,如果列表中的故障原因无法实现故障的排除,则需按照上述步骤重新寻找可能的故障发生原因,直到获得最终的故障发生原因,并依次显示故障传播路径集合  $R$  中所有库所,至此故障诊断过程结束。

4 结束语

本文在有色 Petri 网基本理论的基础上,结合故障诊断的实际,对有色 Petri 网的关联矩阵和状态方程进行改进,将库所作为故障发生和诊断的主要依据,同时动态地更新故障的传播

路径,从而形成了新的基于库所的有色 Petri 网故障诊断模型,并将其应用于船舶电站的故障诊断过程中,实例验证了算法的实用性和高效性。

参考文献:

[1] 魏晓宾,马小平,李亚朋.故障诊断技术综述[J]. 煤矿机电,2009(1):63-65.

[2] 张海艳,夏飞.基于 Petri 网的船舶电力系统故障诊断[J]. 哈尔滨工程大学学报,2008,29(9):933-937.

[3] 刘勇,武昌,陈校平. 随机 Petri 网对装备维修保障的建模与分析[J]. 空军工程大学学报,2007,8(2):43-45.

[4] 王燕平. 基于 Petri 网的舰艇装备故障诊断方法研究[D]. 武汉: 海军工程大学,2009.

[5] PETRI C A. Kommunikation mit automaton[D]. Bonn:Institut Fur Instrumentelle Mathematik Technische Universitat Dannastadt,1962.

[6] 袁崇义. Petri 网原理[M]. 北京:电子工业出版社,1998.

[7] 高方方. 基于着色 Petri 网模型的案例推理研究[D]. 合肥:合肥工业大学,2008.

[8] 王海新. 基于有色 Petri 网舰艇编队 C\_3I 系统的建模与仿真研究[D]. 西安:西安电子科技大学,2005.